U.S. Department of Energy Smart Grid Investment Grant Technical Advisory Group Guidance Document #12

Topic: DOE's Meta Analysis Framework

September 11, 2012

U.S. Department of Energy Smart Grid Investment Grant Technical Advisory Group Guidance Document #12[*]

Topic: DOE's Meta Analysis Framework

September 11, 2011

Objective

This guidance document describes the approach that the Department of Energy (DOE) will utilize in analyzing data produced and reported by Smart Grid Investment Grant (SGIG) recipients who have agreed to undertake a rigorous consumer behavior study (CBS).

Background

The Department of Energy (DOE) set the expectation early on in the SGIG Funding Opportunity Announcement that recipients who undertake a rigorous consumer behavior study would be obliged to collect and report highly granular customer-level consumption and demographic (hereafter referred to as "project") data. DOE provided Guidance Document #10 ("Consumer Behavior Study Data Collection Requirements), which was a description of the totality of data that should be *collected* by SGIG recipients undertaking consumer behavior studies, and Guidance Document #11 ("CBS Data Reporting Process), which described the process by which this CBS data is to be

[*] The following individuals on the Lawrence Berkeley National Laboratory Technical Advisory Group (TAG) drafted and/or provided input and comments on one or more of the U.S. Department of Energy Smart Grid Investment Grant (SGIG) Technical Advisory Group Guidance Documents: Peter Cappers, Andrew Satchwell, Annika Todd and Charles Goldman (LBNL), Karen Herter (Herter Energy Research Solutions, Inc.), Roger Levy (Levy Associates), Theresa Flaim (Energy Resource Economics, LLC), Rich Scheer (Scheer Ventures, LLC), Lisa Schwartz (Regulatory Assistance Project), Richard Feinberg (Purdue University), Catherine Wolfram, Lucas Davis, Meredith Fowlie, and Severin Borenstein (University of California at Berkeley), Miriam Goldberg, Curt Puckett and Roger Wright (KEMA), Ahmad Faruqui, Sanem Sergici, and Ryan Hledik (Brattle Group), Michael Sullivan, Matt Mercurio, Michael Perry, Josh Bode, and Stephen George (Freeman, Sullivan & Company), Mary Sutter and Tami Buhr (Opinion Dynamics). In addition to the TAG members listed above, Bernie Neenan and Chris Holmes of the Electric Power Research Institute also provided comments.

submitted, including a data reporting time frame, a description of the way in which the Project data is to be anonymized and uploaded, and a CBS data dictionary.

This guidance document focuses on a more comprehensive description of the approach that will be used in analyzing project data produced and reported by CBS SGIG recipients and how the reported project data will be used by DOE and its contractor, the Lawrence Berkeley National Laboratory and its subcontractors (LBNL research team) to extract quantitative information for a national audience of policymakers and stakeholders about lessons learned from these studies.

As part of its technical support to DOE Office of Electricity Delivery and Energy Reliability (OE) on SGIG projects, the LBNL research team is planning on undertaking three types of analysis efforts. First, they will examine how customers respond to different treatments (i.e., the effect of the treatments on total energy usage, peak reductions, and peak to off-peak shifting). Second, they will examine the proportion of customers that accept and enroll in different types of programs. Third, they will examine the proportion of customers that are retained throughout the duration of the treatments. For each of these three types of analysis, the LBNL team will analyze treatment level project data, and then will prepare a high-level meta-analysis study that summarizes and synthesizes the treatment level project data. These analytical efforts are discussed in more detail below.

DOE's meta-analysis effort **will not** seek to test or validate the individual recipient's evaluation results.

Customer Response

The LBNL team will examine four specific research questions with project data:

1. **Rates**: What is the effect of time-based rates (e.g., CPP, TOU) on energy usage (including the effect on total energy usage, peak reductions, and peak to off-peak shifting)?
2. **Technology**: What is the effect of enabling technology (e.g., IHDs, web-based education) on energy usage (including effect on total energy usage, peak reductions, and peak to off-peak shifting)?
3. **Customer Characteristics**: What is the effect of time-based rates and enabling technology on energy usage for different customer characteristics? (Customer characteristics studied include income, age, medical needs, education level, housing type, ownership, and energy usage.)
4. **Effect over Time**: What is the effect of time-based rates and enabling technology on energy usage over time?

Analysis

The randomly assigned treatment and control groups in each CBS project will first be validated to ensure that they are equivalent in terms of observable characteristics (including on and off peak energy usage and customer characteristics). Each research question will then be analyzed using a panel data regression technique that compares the difference in the change in energy usage from the pre-treatment time period to the post-treatment time period between the treatment and control group (i.e., the regression will include customer specific fixed effects; one form of a “difference-in-difference” technique).

Because all of the treatments being considered are based on randomly assigned control and treatment groups, the primary regressions will only include treatment variables and fixed effect variables, and will not include any other control or interaction variables. Secondary regressions will include interaction variables that examine the effect of the treatment for different customer characteristics and during different times of the year. All panel data regressions will use robust standard errors for data clustered at the household level. If necessary, additional regressions may be specified for robustness checks.

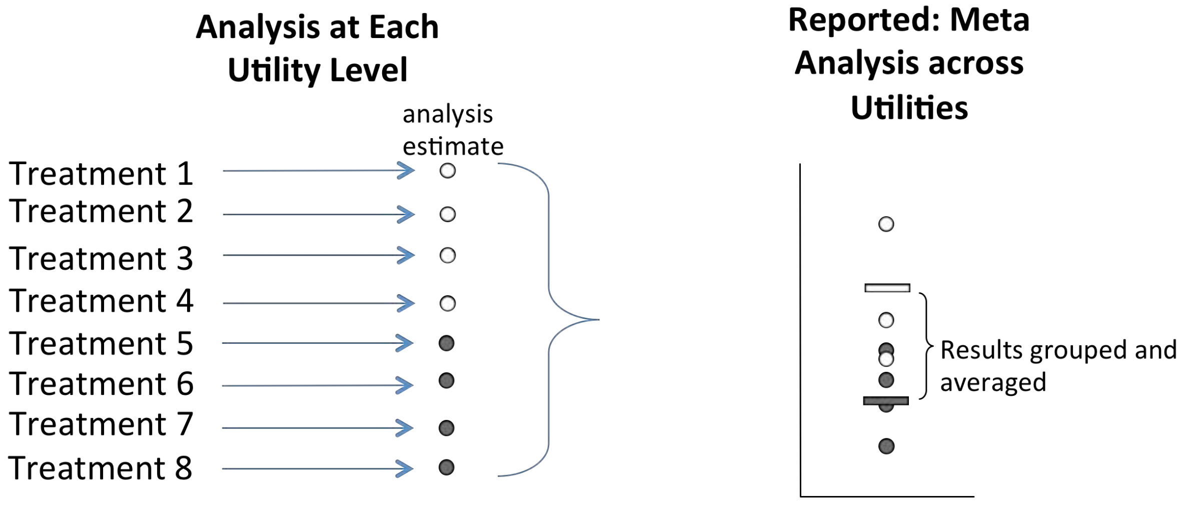
As of the time this guidance document was developed, the following table lists the number of treatments that will be analyzed for each of the four research questions listed above.

Research Question	Treatment being studied	# of Treatments	
		Opt-in	Opt-out
Time-based retail rates	TOU	11	2
	TOU+CPP	14	0
	CPP	3	3
	CPR	0	12
	VPP	5	0
Control and information technology	Education	12	0
	Customer Satisfaction	2	0
	IHD	6	3
	PCT	4	0
	IHD+PCT	17	2
Customer characteristics (7 characteristics)		72	23
Effect over time		72	23

The validation methods and analysis techniques that we will use are specified in the Appendix: “Analysis Methods Guidelines”. In order to answer our research questions about the effect of TOU rates, IHDs, education, and customer satisfaction, we will use the analysis techniques in Appendix Section 2: “Non-Event Treatments”. In order to answer our research questions about the effect of CPP, CPR, and VPP rates, and PCTs, we will use the analysis techniques in Appendix Section 1, “Event-Based Treatments”.

Reporting

Each research question will be analyzed and an effect will be estimated at the treatment level (as described above). The per-treatment analysis estimates will then be grouped and averaged (with an un-weighted average) and reported in an anonymized fashion as a meta analysis, as pictured below. If the results from different treatments are clustered together, they will be segmented first by enrollment method (e.g., opt-in vs. opt-out), then by rate type and/or enabling technology (e.g., CPP, TOU, IHDs, etc.).



Customer Acceptance

The LBNL team will examine five specific research questions with SGIG recipient-provided data:

1. **Enrollment:** What proportion of customers enroll in different types of programs?
2. **Rates**: Do offers of different *rates* affect the proportion of customers that enroll in the program?
3. **Technology:** Do offers of different *enabling technologies* affect the proportion of customers that enroll in the program?
4. **Customer Characteristics:** Do *customer characteristics* affect the proportion of customers that enroll in the program? (Customer characteristics studied include income, age, medical needs, education level, housing type, ownership, and energy usage.)
5. **Technology Installation:** Similar questions for the affect on the proportion of customers that *install the technology* (given that they are assigned to the technology treatment).

Analysis

In order to examine the research questions of interest, the LBNL research team will specify the percentage of customers who enrolled in a program out of the total number of customers who were offered the program in aggregate, and by recruitment method (e.g., the percent of customers who opted-in or who didn't opt-out). In addition, they will also examine a variety of customer characteristics (e.g., the percent of low income customers who joined vs. the percent of high income customers who joined) if there is enough data to support such an analysis. If one proportion is to be compared to another proportion, a z-test of proportions will be performed in order to determine if the difference is statistically significant (e.g., the proportion of customers that enroll in time-of-use program as compared to the proportion that enroll in a critical peak pricing program).

Reporting

Each research question will be analyzed at the program enrollment level. Results from different treatments will be reported in an anonymous way, and may be clustered and averaged in order to determine an effect across treatments and utilities. If the results from different treatments are clustered together, they will be segmented first by enrollment method (e.g., opt-in vs. opt-out), then by rate type and/or enabling technology (e.g., CPP, TOU, IHDs, etc.), and subsequently other customer characteristic information, where applicable.

Customer Retention

The LBNL team will examine five specific research questions with SGIG recipient-provided data:

1. **Retention:** What proportion of people drop out of different types of programs?
2. **Rates:** Do offers of different *rates* affect the proportion of customers that drop out?
3. **Technology:** Do offers of different *enabling technologies* affect the proportion of customers that drop out?
4. **Customer Characteristics:** Do *customer characteristics* affect the proportion of customers that drop out? (Customer characteristics studied include income, age, medical needs, education level, housing type, ownership, and energy usage.)

Analysis

In order to examine the research questions of interest, the LBNL research team will specify the percentage of customers who dropped out of a program out of the total number of customers who were enrolled in the program in aggregate, and by recruitment method (e.g., the percent of customers who opted-in or who didn’t opt-out). In addition, they will also examine a variety of customer characteristics (e.g., the percent of low income customers who dropped out vs. the percent of high income customers who dropped out) if there is enough data to support such an analysis. If one percentage is to be compared to another percentage, a z-test of proportions will be performed in order to determine if the difference is statistically significant (e.g., the proportion of customers that drop out of a time-of-use program as compared to the proportion that drop out of a critical peak pricing program).

Reporting

Each research question will be analyzed at the treatment level. Results from different treatments will be reported in an anonymous way, and may be clustered and averaged in order to determine an effect across treatments and utilities. If the results from different treatments are clustered together, they will be segmented first by enrollment method (e.g., opt-in vs. opt-out), then by rate type and/or enabling technology (e.g., CPP, TOU, IHDs, etc.), and subsequently other customer characteristic information, where applicable.

Appendix: Analysis Methods Guidelines

Table of Contents

Introduction: Statistical Methods for Estimating Effects

This appendix discusses the analysis methods appropriate for estimating the effect of time-based rates, enabling technology, and other treatments on electricity usage in an experimental setting. The LBNL research team will use the guidelines described below to do their analyses. These analysis methods are focused on obtaining accurate after-the-fact estimates of load impacts for the particular case at hand, using as few assumptions as possible.

The terms treatment effect and load impact will be used interchangeably below. Both refer to estimating the effect or impact of rates, technology, or other treatments on energy usage during specific time periods of interest (e.g., during events, peak hours or overall). We will focus on a basic set of analyses that address the following fundamental questions:

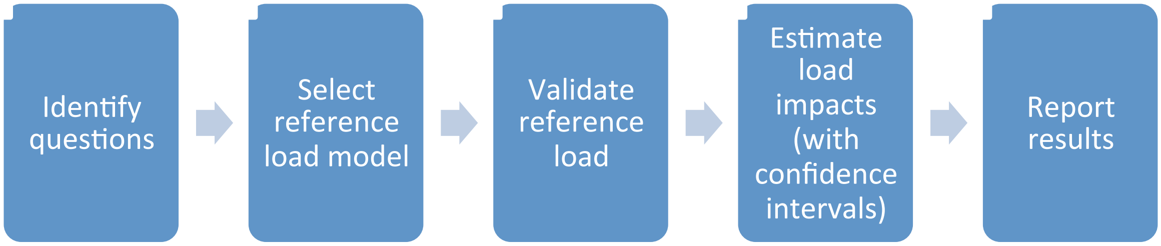
- What was the average load impact of the relevant treatments at specific times of interest?
- How accurate are the estimates?

The regressions that address the first question are all in the form of simple difference-in-difference regressions. These analyses may be expanded to address more detailed questions or perhaps to increase estimation efficiency or precision. However, the analyses should always start by performing and reporting the results of the simple regressions. In this way, a common baseline for comparison can be developed across studies. These results can then provide useful context and corroboration of more complex models used on the same data.

The basic steps for any analysis discussed in this section are to:

- Identify the questions that the analysis must address. In this case, those questions will be restricted to estimating load impacts of particular treatments.
- Select the best possible reference load model (or models), given the data available. The reference load – also known as the estimated load without demand response or the counterfactual load – is an estimate of what the usage would have been among treatment group customers, had they not been exposed to the treatment. This will be determined by the form of the experimental design that was implemented, the question of interest, and the available data;
- Produce validation of the reference load model. This consists of a demonstration that the model accurately predicts load in the treatment group under conditions similar to those of interest, but where loads are observed. This demonstration of the degree of accuracy is crucial for an outsider to be able to interpret the results;
- Estimate load impacts and confidence intervals using the reference load model; and
- Report results.

Figure 3-1: Basic Steps for Analysis


Identify questions
Select reference load model
Validate reference load
Estimate load impacts (with confidence intervals)
Report results

The last step – reporting – consists of four main pieces:

- A verbal discussion of the reference load model, with an emphasis on what experimental design it is based on, why it should provide accurate reference load estimates and any limitations it has;
- A technical description that informs the reader of exactly which analytical steps were taken to produce reference load estimates. For example, in the case of a regression model, this step would consist of reporting the exact regression specification and what data points were used to fit the regression;
- The load impact estimates themselves and their associated confidence intervals or standard errors; and
- Importantly, the results of the validation exercises described in this document that demonstrate how accurate the reference load model is likely to be.

This document is directed at a broad range of evaluators and utility staff with many different backgrounds. We will presume that readers are familiar with basic statistics and econometrics. Additionally, we will assume that readers have basic knowledge of the theory underlying the designs that they are analyzing. For example, we will not lay out the entire theory behind an RED; we will only specify the particular validation exercises and analyses that should be performed for it. Finally, we will assume that there are certain procedures, such as propensity score matching and bootstrapping, that the reader can learn using standard texts.[1]

We will assume that the main focus of the experiment under evaluation is on residential customers who are weather-sensitive in their usage behavior, when averaged over the population. We believe this is

[1] For a comprehensive book on statistical and econometric techniques, see: Greene, W., *Econometric Analysis*, 7th Ed., Prentice Hall, 2011.. For a technical guide to program evaluation see: Imbens, G. M, and J. M Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86. For a guide to program evaluation and implementation see: Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." *Handbook of Development Economics* 4: 3895–3962.

accurate for much of the residential population of the United States due to climate control loads (e.g. air conditioning).

This appendix is divided into two main sections. The first section focuses on estimating and reporting effects for event-based treatments. These treatments are characterized by the fact that during specified hours of the year – referred to as events – and in the hours immediately surrounding them, effects of the treatment are expected to be greater than at any other time by an order of magnitude or more. This means that the focus will be on estimating the effect of the treatment during these event periods and mainly ignoring all other periods.

The second section focuses on estimating and reporting effects for rates and treatments that are always or almost always in effect after their introduction. These treatments are anticipated to have smaller impacts on usage at any given time than during the event periods of event-based treatments. Because they are always in effect, within-subjects designs cannot work as reference load models. Additionally, because expected effect sizes are smaller, the requirements that a model must satisfy for it to produce useful output are more stringent. For example, if a reference load model for an event-based program produces estimates with suspected bias of up to +/-3% of household load, then that might be sufficient because the effect size is around 20% of household load. In contrast, if an IHD is only expected to have an effect of 3% on usage, then a model with +/-3% of bias is much less useful. Underlying this judgment is that a first step in a cost-effectiveness calculation is a determination of whether each program produces a substantial gross benefit prior to cost considerations. An event-based program can be determined to produce a substantial gross benefit, even in the presence of a bias of 3% of household load. On the other hand, an analysis with a bias that large will leave significant doubt about whether a non-event-based program can produce any gross benefit at all.

The third section briefly describes estimating and reporting effects for treatments that contain both event based as well as non-event based elements (e.g., variable peak pricing, real time pricing). The effects of these types of treatments are estimated by using a combination of the analyses described in the event and non-event based sections.

In both the event and non-event based sections, the described analysis will focus first on producing a valid model of reference load. Particular attention will be paid to determining and transparently demonstrating how accurate a given reference load model is likely to be. Once a reference load is determined, estimating treatment effects is fairly straightforward, and the steps will be mostly similar for both types of treatment.

In each section we will begin by describing the ideal reference load design and resulting analysis, and then we will discuss ways to do useful analyses in increasingly compromised experiments.

1 Event-Based Treatments

As discussed in the introduction, event-based treatments consist of three main types: CPP, CPR and utility load control programs. In each case, the utility determines a set number of hours in which customers will either be given a large price incentive to reduce load or will have their load reduced directly for them by the utility. These periods are referred to as events. In the case of CPP and CPR, customers are usually notified the day before the event occurs; and in the case of load control, customers are not typically notified. Each type of program focuses on reducing usage during peak times when wholesale electricity is expensive. Although these programs may produce net energy savings over the course of day, week or billing cycle, they function primarily as a way to reduce peak demand. We focus on analyzing that aspect of the treatment. We will refer to all event treatments generically as a rate in this section.

Events are usually focused on very hot summer afternoons, although some utilities may also use these programs to reduce peak demand in the winter as well. Much of the discussion below will presume that the focus is on estimating effects for summer days. The same discussion can apply to winter days by focusing on very cold days rather than very hot days. A reference load model for this type of program must consist of a method for estimating what the average usage of the treatment population would have been during the event period(s), had they not been exposed to the treatment. We will focus on the task of estimating the average impact of each event separately. The same methods (with small variations) can be used for different levels of possible granularity, such as estimating the impact of each hour of each event, or estimating the degree of pre- or post-event load shifting.

Event based reference loads, and therefore load impacts, can be estimated in two main ways:

- Treatment-control methods use loads observed during the **same** time period for a **different** group of customers who were not subject to an event;
- Within-subjects methods use loads observed during **different** time periods for the **same** group of customers who were subject to an event.

In most cases, treatment-control methods are preferred due to inaccuracy in within-subjects methods. These inaccuracies arise from the necessity of modeling the effects of differences in weather and other changes in conditions between event periods and non-event periods.

Of treatment-control methods, the best reference load model is one based on an RCT. Under an RCT, a population of customers is deemed eligible to receive the treatment, either because they volunteered to receive it (opt-in) or were chosen to receive it (opt-out), and are then randomly assigned to either receive the treatment or serve as the control group. This method is the best from the standpoint of evaluation because it produces an unbiased estimate with the smallest amount of variance per affected customer.

The next best method to use is an RED. Under this method, a population of customers is still subject to random assignment to treatment and control groups, but the treatment consists of an encouragement to take up the rate. The encouragement induces a correlation between being in the treatment group and being on the rate, which can then be used to produce an unbiased estimate of the effect of the rate on the customers who took up the rate because of the encouragement. This is one reason that an RCT is preferable to an RED—the estimated effect for an RCT applies to a broader population than that for an RED. Also, REDs have greater variance per affected customer than for an RCT because customers in the treatment group have the choice not to take the rate, and customers in the control group may have the choice to take the rate. Customers who do so, referred to as non-compliers, add noise to the analysis without providing observable load under the main condition of interest (i.e., an event). It is important to note that virtually any technology-based treatment will have a substantial number of non-compliers due to technology compatibility issues. These non-compliers may be customers who were assigned to the treatment group of an intended RCT; their presence turns the RCT into an RED. It is likely that any technology-based RCT will effectively become an RCT due to compatibility issues.

An RED may be implemented deliberately from the start, but this is also effectively the design for experiments that start out as RCTs, but where customers are allowed to opt out of the treatment group. See section 1.2 for more detail on this point.

If an RCT or RED was not implemented or was unsuccessfully implemented, reasonably accurate estimates of event impacts can likely be produced using a matching estimator to produce a matched control group of non-participants to provide reference load estimates but may be biased in a way that load impact estimates from an RCT or RED are not. Under this method, the reference load model is still one based on observing the loads of customers not in the treatment group, but the method for determining these control group customers differs from an RCT or RED. Although this method can produce precise estimates of load impacts, particularly if hourly interval data is available to use for the matching procedure, we can never be certain of the amount of selection bias introduced by such a procedure. The unobservable characteristics of the control group may be different from those in the treatment group, thereby making the matched control group customers usage patterns a poor counterfactual for the treatment group customers. This is why this method is ranked below RCTs and REDs in this document.

Finally, if the only interval data available is for the group of customers subject to the rate, then it is still worthwhile to estimate the effect of the event period using a reference load model based on loads observed during non-event periods. This within-subjects design typically relies on a panel model or set of regression models on individual customers to produce estimates of reference load that are adjusted for differences in conditions between event periods and non-event periods. There are two major drawbacks to this approach.

First, the question of interest is, what would load have been during events had the customers never been put on the rate. In order to answer this using a within-subjects design, it’s necessary to assume that there is no change in behavior during non-event periods from what would have occurred had the customer not been exposed to the treatment. If that is a weak assumption, the load impact estimates will be biased. We are not aware of any research that addresses the degree to which customers alter their behavior on non-event days due to an event-based rate, except to a very approximate degree (although this could be estimated if an RCT is implemented).

Second, and equally important, a within-subjects design requires adjusting for different conditions between event days and non-event days. This raises a significant risk of misspecification error. Typically only one summer season's worth of data is available and relevant for these analyses. Customer load responds to many factors that do not repeat or repeat only a few times in such a limited dataset. These factors include:

- Temperature patterns, where it is plausible that load may respond to temperatures as much as 24 hours or more in the past due to heat gain in buildings and other structures;
- Seasonality, where it is likely that customers respond differently to the same weather conditions at different times of year due to habit formation or other seasonal issues. For example, customers may be more likely to think to turn on their AC during mid-summer or they may be more likely to be on vacation during the hottest days of summer. Accurately modeling these effects is unlikely because repeated temperature conditions across different parts of the summer are too rare to draw strong conclusions about the effect of seasonality;
- Day of week effects, where it is certain that loads respond differently on weekdays and weekends, but they also likely respond differently in the middle of the week than at the beginning or end. Loads also may respond differently at times surrounding major holidays or during key sporting events; and
- Rare events that may disrupt customer behavior patterns, such as thunderstorms or black outs.

Datasets available for analysis are rarely rich enough to provide much confidence that these factors can be controlled for highly accurately. An additional problem is that the same dataset must be used to both fit the model and then estimate effects. This puts the analyst at risk of over-fitting the model, and it also means that standard errors estimates from these models cannot be interpreted in the usual way. This type of analysis is usually worth performing when there is no alternative, but its limitations should be recognized and reported. Below we provide a set of validation exercises that demonstrate, at least to some degree, how reliable a given within-subjects model is. The results of these exercises should be reported anytime such a model is used.

1.1 Estimation Using an RCT

The best evaluation method for these treatments is a form of difference-in-difference regression performed on a large population of customers that was subject to an RCT to determine assignment to treatment. Under this design, for every event, there will be a sample of customers that is subject to the event (the treatment group), and a sample of customers that is not subject to the event (the control group). To achieve this best outcome, customers must all be recruited into the program and all must satisfy all eligibility requirements. Then a randomization procedure must be used to determine which customers will be in the treatment group and which will be in the control group.

At times there may seem to be ambiguity about exactly which customers are included or excluded from an RCT. A general rule is that any customer excluded from the randomization procedure must be excluded from the analysis and any customer included in the randomization procedure must be included. If customers have the option to opt-out of the treatment or opt-in to the treatment from the control group, then these customers must still be included in the analysis and retained in the group to which they were initially assigned. If a large percentage of customers opted-out, then it may be useful to analyze the RCT as an RED in order to estimate the effect of the treatment only on those who did not opt-out (also known as the treatment effect on the treated), as discussed in Section 1.2.

Customers that close their accounts could be included or excluded, assuming that this type of attrition affects the treatment group and control group similarly. As a theoretical matter, there may be differences between results that include customers who close their accounts during the study and results that exclude those customers. We expect these differences to be small, but results can be estimated both ways to address this issue.

Assuming RCT implementation has been done successfully, the expected value of the average load in the treatment group at any given time is equal to the expected value in the control group at the same time. This is the conceptual basis for the reference load model based on an RCT. There is still an important validation issue in that the treatment group and control group will have different average loads at any given time due to chance alone. This will introduce inaccuracy into impact estimates and will typically be a larger problem when group sizes are smaller. This type of inaccuracy due to chance alone is captured in the standard error or confidence interval of the estimate. However, the standard error also includes the effect of other inaccuracies in the model and therefore does not provide full information about this source of inaccuracy in particular. Additionally, subtle biases can exist in RCT designs that were unsuccessfully implemented; this type of inaccuracy is less easy to determine and subsequently address.

In Section 4, we propose a validation exercise that at least partially addresses the degree to which both of these inaccuracy issues affect the impact estimates of a given RCT. The same basic validation exercise is appropriate for RED and matched control group designs, so we postpone discussion until after discussion of those methods.

Assuming that there is some pre-treatment interval data available for the entire population of interest, the best method to estimate event load impacts in this context is a form of difference-in-differences regression. We use an example to illustrate the method. Suppose we have the set of days described in Table 3-1. Table 3-1 shows a list of two hot pre-treatment days and two CPP event days, along with relevant data for each. For this example, suppose that hourly load data is available for all customers in the treatment and control groups for May-October. Further, suppose that there were only two CPP events, covering the hours 1-5 PM, and that the four days shown are the hottest of the summer, with the next hottest day having a high temperature of 86°F.

Table 3-1: Hypothetical Situation for Modeling the Effect of a CPP Rate Using an RCT

Date	Day Type	Average Load 1-5 PM (kW)		High Temperature (°F)
		Control	Treatment	
6/1/2012	Pre-treatment	1.05	1.11	99
6/15/2012	Pre-treatment	1.10	1.19	94
7/9/2012	CPP Event	1.43	1.05	100
8/1/2012	CPP Event	2.20	1.62	103

Determining exactly which pre-treatment data to include in the model requires judgment about which, if any, pre-treatment periods occurred under conditions similar to the event periods. The above case is fairly typical in that there is a small number of event days and a small number of useful pre-treatment days. We consider days with temperatures more than 5-10 degrees lower than the event days to be not very useful for adjusting for pre-existing differences between groups. Also, in this example it is likely that even the useful pre-treatment days are fairly different from the event days because they are earlier in the summer season and have somewhat lower temperatures than the event days. Despite this, the pre-treatment days are worth including in the model because they show that there is a noticeable difference in average load between the treatment and control group on hot days. Electric usage for a household is based on many factors that are consistent over the course of the summer, so it is fairly likely that this difference reflects a persistent bias between groups. In that case, the estimate can be improved by eliminating it from the model. Most useful would be a set of pre-treatment days with the same conditions as the events, but this situation is unlikely to occur. A good rule to use is that it is better to use pre-treatment data than to not, but it is better to focus only on the most relevant pre-treatment data.

The regression specification for the example in the table is:

$$load_{it} = a_i + b_1 T_i x I_1 + b_2 T_i x I_2 + b_3 I_1 + b_4 I_2 + u_{it} \qquad (1)$$

Table 3-2: Variables in the Regression

Variable	Description
$load_{it}$	Load in kW for customer *i* at time *t*
a_i	Estimated customer-specific additive constant (frequently referred to as a customer fixed effect)
b_1	Estimated average effect of first CPP event
b_2	Estimated average effect of second CPP event
b_3, b_4	Estimated effect of each CPP time period on treatment group and control group customers
I_1, I_2	Indicator variables equal to one during the first and second CPP events, respectively; equal to one for all customers during those events
T_i	Indicator variable equal to one for treatment group customers, zero otherwise
$T_i x I_1, T_i x I_2$	Indicator variables equal to one for treatment group customers during the first and second CPP events, respectively, zero otherwise
u_{it}	The error of the regression for customer *i* at time *t,* which is likely to be correlated over time within customers

The simplest way to represent time is to use the date alone, with the dependent variable load values equal to the average load for each customer during the hours 1-5 PM on each date. An alternative is to keep load measures at the hourly level, which would allow the regression to be altered to measure hourly impacts and also may reduce the standard errors of the estimates.

Standard errors of the treatment effect should be estimated using the cluster option that is available in most statistical packages as part of the regression function. Applying clustering at the customer level will produce estimates that account for correlation in errors at the customer level over time. Not applying clustering at the customer level can produce estimates that appear to be much more accurate than they actually are.

One way to check for errors or possible underlying problems in the data is to compare the results from the difference-in-differences regression to the results of a difference-in-differences aggregated calculation, which can be done based on the information in Table 3-1 alone. Based on Table 3-1, a difference-in-differences aggregated calculation of b_1 is

$$b_1 = (1.43 - 1.05) - \left(\frac{1.05 + 1.10 - 1.11 - 1.19}{2}\right) = 0.45$$

The first term represents the difference between the treatment and control group during the first CPP event. The second term represents the average difference during the two groups during the pre-treatment periods. A simple estimate of the standard error of the estimate is equal to the square root of the sum of the squares of the standard errors of the average difference between the pre-treatment and event load in the treatment and control groups. However, this method of calculating b_1 is likely to result in an estimate that is less precise the method using the panel data regression model in equation (1) because it aggregates observations over time.

Similarly, a difference-in-differences calculation of b_2 equals 0.65. The regression function will produce identical treatment effect estimates to the regression except in cases where mistakes have been made in implementing one or the other.

The difference-in-difference regression is based on the assumption that pre-treatment data is available, which may not always be true. In that case, it is impossible to adjust for pre-treatment differences between treatment and control groups, but it is possible to estimate load impacts using load in the control group as a reference load, with a regression specification:

$$load_{it} = a + b_1 T_i x I_1 + b_2 T_i x I_2 + u_{it} \qquad (2)$$

Note that here there are no customer-specific constants. In that case, a way to check for errors or possible underlying problems in the data is to compare the results from the regression to the results of subtracting the average load in the treatment group during the event from the average load in the control group during the event, which could be done with the last two rows of Table 3-1 (i.e., excluding the pre-treatment numbers). The standard error of the estimate is equal to the square root of the sum of the squares of the standard errors of the average load in the treatment and control groups, the typical calculation for estimating standard errors of differences.

1.2 Estimation using an RED

An RED is a form of RCT in which a randomized group receives encouragement to take up treatment, but may or may not do so. For example, the encouragement might consist of a marketing effort directed at a randomly-chosen subset of the eligible population for a particular time-based retail rate. By directing the marketing at only a subset of the population, more customers in that subset will sign up for the rate than will do so in the remaining eligible population. This induced difference in rate take-up between groups in the population is the basis for an analysis of an RED design. As discussed in Section 1.1, an RCT in which treatment group customers can opt-out of the treatment or control group customers can opt-in is actually an RED. While it can be analyzed as an RCT, including all of the customers originally assigned to the treatment and control groups (know as an intent-to-treat estimate), it might be beneficial to also analyze it as an RED, especially if more than a few percentage points of either the treatment or control group become non-compliers.[2]

It is important to distinguish here between customers opting out of the treatment group and customers moving away or otherwise dropping out of the utility's billing system. Customers who opt out of the treatment group cause a selection bias problem for an RCT. They must be retained in the study, and considered part of the treatment group even if they are no longer on the rate. This changes the study from an RCT to an RED in which customers who were initially assigned to the treatment group constitute the encouraged group.

[2] This determination is inherently subjective. A good way to deal with this issue when the situation is ambiguous is to analyze the data both as an RCT and as an RED and see whether results change substantively.

There are two fundamental questions that must be addressed regarding the validity of the RED. First, how comparable was the encouraged group to the non-encouraged group? This is directly analogous to the question of how comparable are treatment and control groups in an RCT design. We discuss the best way to address this question in Section 4 below.

Second, how strong was the effect of the encouragement? Theoretically, even a weak encouragement could work for estimating load impacts if sample sizes were large enough, but as a practical matter it must be the case that a large fraction of customers who receive encouragement actually take up the treatment and that only a small fraction of customers who do not receive encouragement do so. This should be reported using a table such as Table 3-3 shown below:

Table 3-3: Percentage of Customers in the Encouraged and Not Encouraged Groups Accepting the Rate

Group	Accept Rate	Refuse Rate
Encouraged	65	35
Not Encouraged	2	98

Table 3-3 shows a reasonably successful RED in which almost two-thirds of encouraged customers took the offered rate and a very small fraction of non-encouraged customers did so. In the case of an RCT where customers are allowed to opt out, the encouraged group consists of the treatment group, and the fraction accepting the rate is the fraction that chooses not to opt out.

To estimate load impacts the most straightforward method is to use the same regression function described for RCTs to first estimate an intent-to-treat effect of the encouragement. In this case the RCT regression, equation (1) in Section 1.1, is altered so that T_ixI_1 and T_ixI_2 are equal to one for any customer in the encouraged group during critical peak event 1 and 2, respectively. This means that the estimated coefficients on those variables represent the effect of a CPP event on customers encouraged to take the rate rather than on customers who are necessarily on the rate.

The effect of the rate, rather than the encouragement, can be estimated by dividing the effect of the encouragement for each event (i.e., b_1 or b_2 from equation (1)) by the difference in the fraction of customers who took up the rate between the encouraged group and the non-encouraged group. For example, in Table 3-3, that fraction equals 0.65-0.02=0.63. This is equivalent to scaling up the encouragement effect for each event by 1/0.63=1.59. Note that this fraction might be different for each CPP event if customers moved in and out of the rate between events.

The standard error of the effect of the rate rather than the encouragement can be developed similarly. Take the standard error of the effect of the encouragement and divide it by the difference in the fraction of customers who took up the rate between the encouraged group and the non-encouraged group. This calculation illustrates how lower take up levels in the encouraged group leads to inflated standard errors in the treatment effect.

Alternatively, the treatment effect can be estimated directly using an instrumental variables approach. Taking as an example the hypothetical CPP situation from Table 3-1, the first stage regressions have as their dependent variables indicators for being on the rate during each CPP event and as their independent variables a constant, an indicator for being in the encouraged group and indicator variables for each CPP period. Note that the first stage is estimated over all time periods, including pre-treatment time periods. The dependent variables in the first stage regressions will equal zero for all customers during all pre-treatment time periods. There are two variables to be instrumented for: being on the CPP rate during the first CPP event and being on the CPP rate during the second CPP event. Consequently we need at least two instruments. There are exactly two instruments: being in the encouraged group during the first CPP event and being in the encouraged group on the CPP rate during the second CPP event.

The second stage is then the same regression specification as in equation (1), but using the predicted values from the first stage instead of T_ixI_1 and T_ixI_2. This system of equations is shown below[3]:

$$T_i x I_1 = a_{1i} + b_5 I_E x I_1 + b_6 I_E x I_2 + b_7 I_1 + b_8 I_2 + \varepsilon_{it} \qquad (2)$$

$$T_i x I_2 = a_{2i} + b_9 I_E x I_1 + b_{10} I_E x I_2 + b_{11} I_1 + b_{12} I_2 + \epsilon_{it} \qquad (3)$$

$$load_{it} = a_i + b_1 \widehat{T_\imath x} I_1 + b_2 \widehat{T_\imath x} I_2 + b_3 I_1 + b_4 I_2 + u_{it} \qquad (4)$$

[3] We used the specification shown because they are the easiest to implement in a statistical package such as Stata. Because the variables from the second CPP event aren't an instrument for being on the rate during the first CPP event, the resulting coefficients for the variables related to the second event will be zero. An alternate specification might include only variables from the first CPP event in one first stage regression and variables from the second CPP event in another first stage regression.

Table 3-4: Variables in the Regression

Variable	Description
$load_{it}$	Load in kW for customer *i* at time *t*
a_i, a_{1i}, a_{2i}	Estimated customer-specific additive constants
b_1	Estimated average effect of first CPP event on customers who are on CPP
b_2	Estimated average effect of second CPP event on customers who are on CPP
b_3, b_4	Estimated effect of CPP time periods on treatment group and control group customers
b_5, b_6, b_9, b_{10}	Estimated effect of the encouragement on CPP take-up during each CPP event; note that b_6 and b_9 should be very close to zero
b_7, b_8, b_{11}, b_{12}	Estimated effect of CPP time periods in the first stage regressions
I_1, I_2	Indicator variable equal to one during the first and second CPP events, respectively; equal to one for all customers during those events
I_E	Indicator variable equal to one for customers in the encouraged group, zero otherwise
T_i	Indicator variable equal to one for treatment group customers, zero otherwise
$T_i x I_1, T_i x I_2$	Indicator variables equal to one for treatment group customers during the first and second CPP events, respectively, zero otherwise
$\widehat{T_i x I_1}, \widehat{T_i x I_2}$	Fitted values from the first stage regressions
u_{it}	The error of the regression for customer *i* at time *t,* which is likely to be correlated over time within customers
$\varepsilon_{it}, \epsilon_{it}$	Error terms in the first stage regressions

This two-stage process might have to be done more than once if the fraction of customers on the rate changed significantly between events. In that case, the same specification holds, but only events where the fraction of customers on the rate stayed nearly constant should be included in the same estimation.

The two-stage process should produce treatment effect estimates identical to those obtained through the first method of running the regression on the encouraged variable and dividing by the difference in the take-up rate between encouraged and non-encouraged customers. The two-stage process will produce slightly different standard errors than the first method, but they are unlikely to be materially different.

1.3 Estimation Using a Matched Control Group

If an RCT or RED was not implemented, or was implemented unsuccessfully, then the next best method for load impact estimation is to identify a matched control group and then perform the same difference-in-difference regression analysis described in Section 1.1. This is inferior to the RCT or RED designs for two reasons.

First, there will be a very limited set of observable variables to use for matching customers. This means that there will probably be noticeable biases between the groups, regardless of the matching procedure that is used.

Second, regardless of how well the matching procedure appears to work, we can never rule out the possibility of unobservable differences between groups producing bias in the results. In the case of event-based programs where the underlying rate in the treatment group is the customer's default rate, these potential biases can be bracketed as likely to be small if we observe only small differences between the treatment group and the matched control group on hot non-event days during the treatment period.

The matching process uses propensity score matching to identify customers not subject to the treatment who have similar characteristics to the customers who are subject to the treatment. There are several similar methods that can be used to control for selection. For this method, because the matching process and the resulting panel regression are modular, this process allows for the entire second half of the analysis to be identical to the analysis for an RCT. Other methods include the correction for selection in the regression itself.

This matching process could be used if no control group was ever chosen by selecting customers from the utility's broader population. It could also be used to construct a better control group from one that was corrupted during implementation.

In each case, propensity score matching is based on the following observable variables in order of priority, with the understanding that not all of them will be available:

- Hourly usage at the customer level during the afternoon and evening on hot, pre-treatment days;
- Hourly usage at the customer level during the afternoon and evening on hot, non-event, post-treatment days if pre-treatment data is not available and the event-based rate is the only treatment the customer is subject to (i.e. the underlying rate is not TOU and the customer has not been provided with an IHD or PCT);
- Customer-level demographic and location information, such as size of house, income, number of people in the household, age, zip code; and
- Aggregate level demographic information, such as income, size of house, age, based on Census block group data.

Having performed the propensity score match to produce a control group, the regression analysis to measure load impacts is the same as that described in Section 1.1.

1.4 Within-Subjects Methods

It may be impossible to use treatment-control methods in some cases. For example, it may be the case that smart meter interval data is only available for customers exposed to the treatment. In the case of event-based treatments, a reference load model and an estimate of the effect of the treatment can still be produced using load observed within the treatment group during non-event periods.

As mentioned above, the interpretation of the reference load estimate for an event in this case is that it represents an estimate of what load would have been in the treatment group had an event not been called. The validity of the within-subjects method relies on two assumptions. First, it assumes that there is a set of non-event days that have sufficiently comparable characteristics to the event days so that it is reasonable to assume that the only difference in customer behavior during event days and the comparable non-event days is that they experienced the treatment in effect during event days (e.g., the comparable non-event days have similar temperatures and daylight hours to the event days). Second, if the underlying rate is the default rate in the population and if it seems reasonable to assume that customers do not alter their non-event behavior very much due to being on the rate, then this estimate can stand in as a decent approximation of what load would have been if the treatment group had never been on the rate. This may frequently be the case.

Within-subjects models consist of either panel regressions on only the treatment group or regressions run individually for each customer. In each case, event impacts are measured using indicator variables for event periods. We will not go into the details of this method here, as there are many examples in the industry of these types of analysis. We will only provide these guidelines:

- There is little value to including in the dataset non-event days where temperatures are far different than event temperatures. Given that CPP days are often the hottest days of the summer, a typical problem in the model is lack of relevant non-event days for modeling;
- In determining what variables to try to include and what conditions to try to control for, it is best to keep expectations modest. The amount of independent variation in weather at different times of day over the course of a summer is usually low. The same is true of variation in seasonal conditions and day-of-week conditions independent of weather;
- The accuracy of any model should be assessed using an out-of-sample testing regime in which several non-event days with event-like conditions are withheld from the model during fitting. Load predictions from the model can then be compared to actual load on these days to assess predictive accuracy. The result of this exercise should be displayed in graphs similar to Figures 3-1 and 3-2. This exercise also limits the potential for over-fitting the model to the data by adding lots of variables; and
- Over-fitting to the out-of-sample days is still possible. Whether a model is being over-fit requires some judgment. The question for any given variable is, is it really plausible that there is enough variation in this variable, independent of all the other variables, for this coefficient to be well-measured? For example, if we think that morning temperature in particular is a useful predictor of CPP impacts in addition to daily average temperature, we need to determine whether the data provides significant variation in morning temperature for the relevant range of daily average temperature.

2 Non-Event Treatments

Non-event treatments consist of several types: TOU rates, IHDs, PCTs, educational efforts, and combinations of these. We will refer to all three generically as treatments in this section. In each case, customers are exposed to a treatment that provides a motivation to change usage behavior, either through incentives, by providing more information about usage behavior, or by providing tools that more easily allow changes in usage behavior. In each case, the treatment is always present after its onset rather than only being present during certain critical events. For these types of treatments it might be reasonable to hypothesize that in addition to leading to a reduction during certain, targeted times of the day, the treatment might also lead to an overall reduction in energy use and also a shifting of usage from targeted times to non-targeted times of day. We will discuss analyzing both of these effects.

In addition to the fact that these treatments are always present, there is another important difference between the event and non-event based treatments that directly affects our ability to measure impacts of non-event based-treatments. The effect of non-event-based treatments on usage at any given time is likely to be smaller than that of an event-based treatment during an event. This means that as a practical matter there are more severe analytical requirements for measuring the effect of non-event-based treatments well enough to establish their value. The price signal from most TOU rates is quite mild and effects on usage are usually no larger than about 5% during any given hour. Currently the effects of IHDs, PCTs or education on usage have not been rigorously tested, but those effects are also likely to be smaller than the impacts of event-based programs during an event. The energy conservation effects of these programs are also likely to be modest – in the range of 0-5% of usage.

Given these characteristics, it is unlikely that practically useful estimates of the effects of these treatments can be developed using any design other than an RCT or an RED with a low rate of non-compliers. Other methods tend to have noise in the results at least as large as the effects being measured. For that reason, we limit our discussion to estimating energy use effects and shifting effects using these two designs.

We discuss two main types of estimation for non-event-based treatments: demand-shifting and energy conservation. We use demand-shifting to refer to any change in usage during particular hours, whether or not that usage change is made up for during other hours. It may or may not be, and an estimation scheme that covers all hours of the day can answer that question.

2.1 Estimation Using an RCT

The best method to estimate energy usage effects and shifting effects for these treatments is a form of difference-in-difference regression performed on a large population of customers that was subject to an RCT to determine assignment to treatment. Under this design, there is a sample of customers subject to the treatment (the treatment group), and a sample of customers that is not subject to the treatment (the control group). To achieve this best outcome, customers must all be recruited into the program and all must satisfy all eligibility requirements. Then a randomization procedure must be used to determine which customers will be in which group.

At times there may seem to be ambiguity about exactly which customers are included or excluded from an RCT. A general rule is that any customer excluded from the randomization procedure must be excluded from the analysis and any customer included in the randomization procedure must be included. If customers have the option to opt-out of the treatment or opt-in to the treatment from the control group, then these customers must still be included in the groups to which they were initially assigned. If a large percentage of customers opted-out, then it may be useful to analyze the RCT as an RED in order to estimate the effect of the treatment only on those who did not opt-out (also known as the treatment effect on the treated), as discussed in Section 2.2.

Customers that close their accounts could be included or excluded, assuming that this type of attrition affects the treatment group and control group similarly. As a theoretical matter, there may be differences between results that include customers who close their accounts during the study and results that exclude those customers. We expect these differences to be small, but results can be estimated both ways to address this issue.

Assuming RCT implementation has been done successfully, the expected value of the average load in the treatment group at any time is equal to the expected value in the control group at the same time. This is the conceptual basis for the reference load model based on an RCT. There is still an important validation issue in that the treatment group and control group will have different average loads at any particular time due to chance alone. This will introduce inaccuracy into impact estimates and will typically be a larger problem when group sizes are smaller. This type of inaccuracy due to chance alone is captured in the standard error or confidence interval of the estimate. However, the standard error also includes the effect of other inaccuracies in the model and therefore does not provide full information about this source of inaccuracy in particular. Additionally, subtle biases can exist in supposedly well-implemented RCT designs.

In section 4 we propose a validation exercise that at least partially addresses the degree to which both of these issues affect a given RCT. The same basic validation exercise is appropriate for RED designs, so we postpone discussion until after discussion of those methods.

2.2 Estimating Demand Shifting using an RCT

Estimating the demand-shifting effect of these rates based on an RCT is best performed using a difference-in-differences regression, similar to the one for event-based programs in Section 1.1. It may be of interest to measure usage shifting behavior for any particular hour of the day, or a particular block of hours. Consider the example where our interest is to estimate the degree to which a treatment causes customers to shift usage away from the time period 1-5 PM on weekdays.

The primary analysis would consist of estimating an effect that is specific to the days and/or months that are of interest, and separating the analysis as such (e.g., if the target of the treatment is reduction during weekdays, then weekdays should be separated from weekends; likewise if summer is the target it should be separated from other seasons). As a secondary analysis, a further separation of days (or an inclusion of interaction variables for those categories) may provide insight into the effectiveness of the treatment. Specifically, dividing days into at least two categories based on weather conditions may be useful if customers may have more scope to shift load at times when it is very hot or very cold.

In this example, the regression to estimate the effect of the treatment under normal conditions during 1-5 PM on weekdays in the summer is:

$$load_{it} = a_i + b_1 T_i x I_1 + b_2 I_1 + u_{it} \qquad (5)$$

Table 3-5: Variables in the Regression

Variable	Description
$load_{it}$	Load in kWh for customer *i* at time *t,* only including weekdays in the summer under normal conditions
a_i	Estimated customer-specific additive constant
b_1	Estimated average effect of the treatment
b_2	Estimated effect of the treatment period for all treatment group and control group customers
I_1	Indicator variable equal to one during treatment periods, zero otherwise
$T_i x I_1$	Indicator variable equal to one during the treatment period for treatment group customers, zero otherwise
u_{it}	The error of the regression for customer *i* at time *t,* which is likely to be correlated over time within customers

The load data that should be included in the regression are loads measured at the customer level during the periods of interest in that specific regression: in this example, summer weekdays from 1-5 PM. Both the treatment and pre-treatment periods should be included. Separate regressions can be run for different combinations of day type, season and conditions. Similarly, the regressions can be run for any time of day to determine the average impact at that time (e.g., the regression could be also run for every hour except for 1-5 PM in order to estimate the degree to which usage shifted from peak to off peak). It is also possible to use interactions of indicator variables to produce these results all at once, but it is sufficient to use separate regressions so we do not discuss that further.

As an example of the secondary analysis mentioned above, suppose the area of interest is in the Midwest and we divide all days into the categories "normal" and "extreme", where "normal" is defined as having high temperatures in the range 25-85°F and extreme is defined as temperatures outside that range. This categorization separates summer days between those that are mild and those that are very hot, and winter days between those that are mild and those that are very cold. The specification in (5) could then be used separately for days in each category.

Standard errors of the treatment effect should be estimated using the cluster option that is available in most statistical packages as part of the regression function. Applying clustering at the customer level will produce estimates that account for correlation in errors at the customer level over time.

2.3 Estimating Energy Conservation using an RCT

Estimating energy conservation effects of one of these treatments does not require smart meter interval data; it can be performed using monthly billing data alone (although using smart meter interval data may increase the precision of the estimate). Estimating the overall energy savings associated with such a treatment is done using a specification similar to equation (5). Suppose that we are estimating the energy savings associated with an IHD that has been in place for a treatment group for a year and that we have one year of pre-treatment monthly billing data for all customers in the treatment group and the control group. To estimate the overall average energy savings associated with the IHD during the year-long experiment, we would regress monthly usage for each customer for each month onto a customer-specific constant, an indicator equal to one during the treatment period and zero otherwise and an indicator equal to one for customers in the treatment group during the treatment period and zero otherwise. This specification is:

$$load_{im} = a_i + b_1 T_i x I_1 + b_2 I_1 + u_{im} \quad (6)$$

Table 3-6: Variables in the Regression

Variable	Description
$load_{im}$	Monthly usage in kWh for customer *i* during month *m,* only including weekdays in the summer under normal conditions
a_i	Estimated customer-specific additive constant
b_1	Estimated average effect of the treatment
b_2	Estimated effect of the treatment period on treatment and control group customers
I_1	Indicator variable equal to one during the treatment period
$T_i x I_1$	Indicator variable equal to one during the treatment period for treatment group customers, zero otherwise
u_{im}	The error of the regression for customer *i* during month *m,* which is likely to be correlated over time within customers

A common variation on this specification includes indicator variables for each month-year combination rather than a single indicator variable for the treatment period. Additionally, this specification can be adapted to include different estimated effects for each month. Standard errors of the treatment effect should be estimated using the cluster option that is available in most statistical packages as part of the regression function. Applying clustering at the customer level will produce estimates that account for correlation in errors at the customer level over time.

Rather than estimating the model using absolute load values, there might be interest in using the logarithm of load as the dependent variable in order to directly estimate percentage reductions in energy usage. While there is some value in that exercise,[4] estimating the model on raw load values and then converting the estimated effect to a percentage of observed usage is preferable. Estimating the effect using the logarithm of usage tends to equalize the effect of customers with different levels of overall usage. If large customers tend to reduce usage much more than small customers as a fraction of their overall usage, then this will tend to understate the energy conservation effect in the population. Keeping the dependent variable linear avoids this problem. An alternative would be to use the logarithm of usage, but to fit separate treatment effects for customers based on their overall level of usage.

[4] In particular, the time-based indicator variables and customer-specific indicator variables tend to explain more of the variation in usage in a logarithm model than in a linear model.

One issue that has to be dealt with is that most utility billing is done on monthly cycles that have bill dates that differ across customers. This means that some approximation has to be used to assign customer usage values to particular months. A simple method is to assign to a particular month all bills that were finalized prior to the middle of the next month and to assign to the next month all bills that were finalized after that. The month during which each customer was enrolled in the treatment could be excluded from the analysis because it will constitute a partial treatment month. A more complicated method that may be more accurate is to assign each day a usage value based on the monthly bill divided by the number of days in that billing cycle. Then, each daily value can be assigned to the month in which it occurs. For example if a customer used 600 kWh in a billing cycle from April 15-May 14, then the daily value for each of those days would be 600 kWh/30 days=20 kWh/day. Then that value would be assigned to the last half of the days in April and the first half of the days in May. Again, the month each customer enrolled on the treatment could be excluded.

2.4 Estimation Using an RED

The basic logic of the RED design is identical for event-based and non-event-based treatments. We refer the reader to Section 1.2 for a brief conceptual description of REDs and how to use them for estimation. We also refer the reader to Section 4 for a description of the validation exercises for REDs. In this section we concentrate specifically on how to specify the regression model for an RED. As in the case of an RCT for a non-event-based program, we discuss two main types of estimation: demand-shifting and energy conservation.

2.5 Estimating Demand-Shifting using an RED

Just as in Section 1.2, there are two main ways to develop load impact estimates and standard error estimates for an RED design. The most straightforward method to develop load impacts is to use the same regression function described for RCTs to first estimate an intent-to-treat effect of the encouragement. In this case, the RCT regression, equation (5) in Section 2.2, is altered so that T is equal to one for any customer in the encouraged group during the treatment period. This means that the estimated coefficient on that variable represents the effect of the treatment on customers encouraged to take the treatment rather than on customers who necessarily took the treatment.

The effect of the treatment, rather than the encouragement, can be estimated by dividing the effect of the encouragement (i.e., b_1 from equation (5)) by the difference in the fraction of customers who took up the rate between the encouraged group and the non-encouraged group. For example, looking back to the example in Table 3-3, that fraction equals 0.65-0.02=0.63. This is equivalent to scaling up the encouragement effect by 1/0.63=1.59.

As discussed in section 3.1.2, to develop standard error estimates for the load impact estimate, the standard error estimates of the encouragement effect can also be multiplied by the difference in the fraction of customers who took up the rate between the encouraged group and the non-encouraged group.

As an alternative to scaling the regression results of the encouragement, we can specify the regression in terms of instrumental variables, with the encouragement during the treatment period acting as an instrument for the treatment during the treatment period. Taking as an example the hypothetical situation in equation (5), but supposing that an RED was implemented instead of an RCT, the first stage regression has as its dependent variable an indicator for having the treatment during the treatment period and as its independent variables a constant, an indicator for being in the encouraged group and an indicator variable for the treatment period. Note that the first stage is estimated over all time periods, including pre-treatment time periods. The dependent variable in the first stage regression will equal zero for all customers during all pre-treatment time periods. The second stage is then the same regression specification as in equation (5), but using the predicted values from the first stage in place of $T_i x I_1$. This system of equations is shown below:

$$T_i x I_1 = a_{1i} + b_3 I_E x I_1 + b_4 I_1 + \varepsilon_{it} \qquad (7)$$

$$load_{it} = a_i + b_1 \widehat{T_i x I_1} + b_2 I_1 + u_{it} \qquad (8)$$

Table 3-7: Variables in the Regression

Variable	Description
$load_{it}$	Load in kWh for customer *i* at time *t,* only including weekdays in the summer under normal conditions
a_i, a_{1i}	Estimated customer-specific additive constants
b_1	Estimated average effect of the treatment
b_2	Estimated effect of the treatment period on all customers
b_3	Estimated effect of encouragement on treatment take-up
b_4	Estimated effect of the treatment period in the first stage
I_1	Indicator variable equal to one during the treatment period, zero otherwise
I_E	Indicator variable equal to one for customers in the encouraged group, zero otherwise
T_i	Indicator variable equal to one for treatment group customers, zero otherwise
$T_i x I_1$	Indicator variable equal to one during the treatment period for treatment group customers, zero otherwise
$\widehat{T_i x I_1}$	Fitted values from equation (7)
u_{it}	The error of the regression for customer *i* at time *t,* which is likely to be correlated over time within customers
ε_{it}	The error in the first stage regression

If the fraction of customers on the rate in the encouraged and non-encouraged groups changes significantly during the study period, then the above analysis should be repeated separately for time periods of relatively stable treatment rates.

Just as in the RCT case, standard errors should be estimated using the cluster option that is available in most statistical packages as part of the regression function. Applying clustering at the customer level will produce estimates that account for correlation in errors at the customer level over time.

The two-stage process should produce treatment effect estimates identical to those obtained through the first method of running the RCT regression on the encouraged variable and dividing by the difference in the take-up rate between encouraged and non-encouraged customers. The two-stage process will produce slightly different standard errors than the first method, but they are unlikely to be materially different.

2.6 Estimating Energy Conservation using an RED

The same basic principles discussed in Section 2.5 above apply to using an RED to estimate energy conservation rather than demand shifting. First, to develop an estimate of the intent-to-treat effect, an RCT regression can be used with a variable for being in the encouraged group used in place of the treatment group variable. To estimate the effect of the treatment on customers who took the treatment, that estimate can then be scaled by the difference in the fraction of customers who took up the rate between the encouraged group and the non-encouraged group. See Section 2.5 for an example.

Again, standard error estimates can be developed by multiplying the standard error estimates for the encouraged variable in the RCT regression on encouragement by the difference in the fraction of customers who took up the rate between the encouraged group and the non-encouraged group.

Again, as in Section 2.5 and extending the example by supposing the situation in equation (6) is implemented as an RED rather than an RCT, we can also specify the RED using instrumental variables. The first stage regression has as its dependent variable an indicator for having the treatment and as its independent variables a constant, an indicator for being in the encouraged group during the treatment period and an indicator equal to one during the treatment period and zero otherwise. The second stage regression is the same as equation (6), but with the fitted values from the first stage used in place of the treatment variable.

$$T_i x I_1 = a_{1i} + b_3 I_E x I_1 + b_4 I_1 + \varepsilon_{im} \qquad (9)$$

$$load_{im} = a_i + b_1 \widehat{T_i x I_1} + b_2 I_1 + u_{im} \qquad (10)$$

Table 3-8: Variables in the Regression

Variable	Description
$load_{im}$	Monthly usage in kWh for customer *i* during month *m,* only including weekdays in the summer under normal conditions
a_i, a_{1i}	Estimated customer-specific additive constant
b_1	Estimated average effect of the treatment
b_2	Estimated effect of the treatment period on all customers
b_3	Estimated effect of encouragement on treatment take-up
b_4	Estimated effect in the first stage regression of the treatment period on all customers
T_i	Indicator variable equal to one for treatment group customers, zero otherwise
I_1	Indicator variable equal to one during the treatment period, zero otherwise
I_E	Indicator variable equal to one for customers in the encouraged group, zero otherwise
$T_i x I_1$	Indicator variable equal to one for treatment group customers during the treatment period, zero otherwise
$\widehat{T_i x I_1}$	Fitted values from equation (9)
u_{im}	The error of the regression for customer *i* during month *m,* which is likely to be correlated over time within customers
ε_{im}	The error of the first-stage regression for customer *i* during month *m*

As in the case of equation (6), a common variation would include indicator variables for each month-year combination. Standard errors should be estimated using the cluster option that is available in most statistical packages as part of the regression function. Applying clustering at the customer level will produce estimates that account for correlation in errors at the customer level over time.

The two-stage process should produce treatment effect estimates identical to those obtained through the first method of running the RCT regression on the encouraged variable and dividing by the difference in the take-up rate between encouraged and non-encouraged customers. The two-stage process will produce slightly different standard errors than the first method, but they are unlikely to be materially different.

3 VPP and RTP

VPP and RTP rates have some characteristics of both event-based and non-event based treatments. Like non-event-based treatments, they are always in effect; and like event-based treatments, they present customers with particularly high prices at times that are not known far in advance. That they are always in effect restricts the type of analysis that can be done to examine their effects to those based on the use of a control group. For both VPP and RTP, the same basic analyses for measuring shifting and conservation associated with TOU rates in Section 2 can be used. If there are periods of high prices that are of particular interest to analyze separately from other periods, then the control group methods for event-based programs in Section 1 can be used. In each case, the validation exercise described in Section 4 should also be performed.

4 Validation for RCTs, REDs and Matched Control Groups

The common element of the treatment-control methods discussed in the context of both event and non-event based treatments is that the reference load estimate is determined by load observed among a group of customers chosen to be comparable to the customers in the treatment (or encouraged) groups. Therefore, to understand how accurate load impact estimates are likely to be, it is crucial to understand how similar those groups are. The groups that must be similar for each experimental set up and analysis type are:

- The treatment and control groups in an RCT;
- The encouraged and non-encouraged groups in an RED; and
- The treatment and matched control groups in a propensity score matching analysis.

This section describes two simple ways of illustrating this comparability that can be applied to any of these situations. In order for an outsider to be able to properly interpret the results of an experiment or quasi-experiment, these are crucial analyses to include.

The close comparability of the groups is necessary because there are severe limitations to what can be corrected for using regression or other statistical methods. Customer-based fixed effects, such as those in the models in Section 1.1 or 2.1, are useful in removing unexplained variation from a model, but they can only correct for differences between groups that do not change over time. More complicated modeling to account for non-comparability, such as specifying load as a function of recent temperatures and daily and seasonal patterns is subject to misspecification and lack of data for modeling. Such modeling can certainly improve estimates, but the degree of accuracy becomes much harder to assess. Therefore, load impact estimates are much more reliable if the control group is quite similar to the treatment group to begin with.

There are two main types of data used to assess comparability between groups: demographic data; and usage data, preferably at the hourly level. The number of customers in each group who close their accounts during the experiment should also be tracked and reported.

An example of how to demonstrate the comparability of groups using demographic data is Table 3-5.[5] Table 3-5 shows the results of an actual propensity score match used to estimate the impact of Pacific Gas & Electric's (PG&E's) SmartRate, a residential CPP rate. The table shows the fraction of the treatment group ("SmartRate Population") and the fraction of a matched control group that is located within each of seven local capacity areas, an important geographical characteristic. It also shows average monthly usage for two summer months for both groups and the fraction of customers in each group that is on PG&E's CARE tariff, an underlying rate for low-income customers. The table also shows t-statistics and p-values for the differences between the groups. Depending on the audience, a normalized difference between the two groups for each characteristic may be a more appropriate method

[5] Taken from FSC's 2011 evaluation of PG&E's SmartRate CPP program.

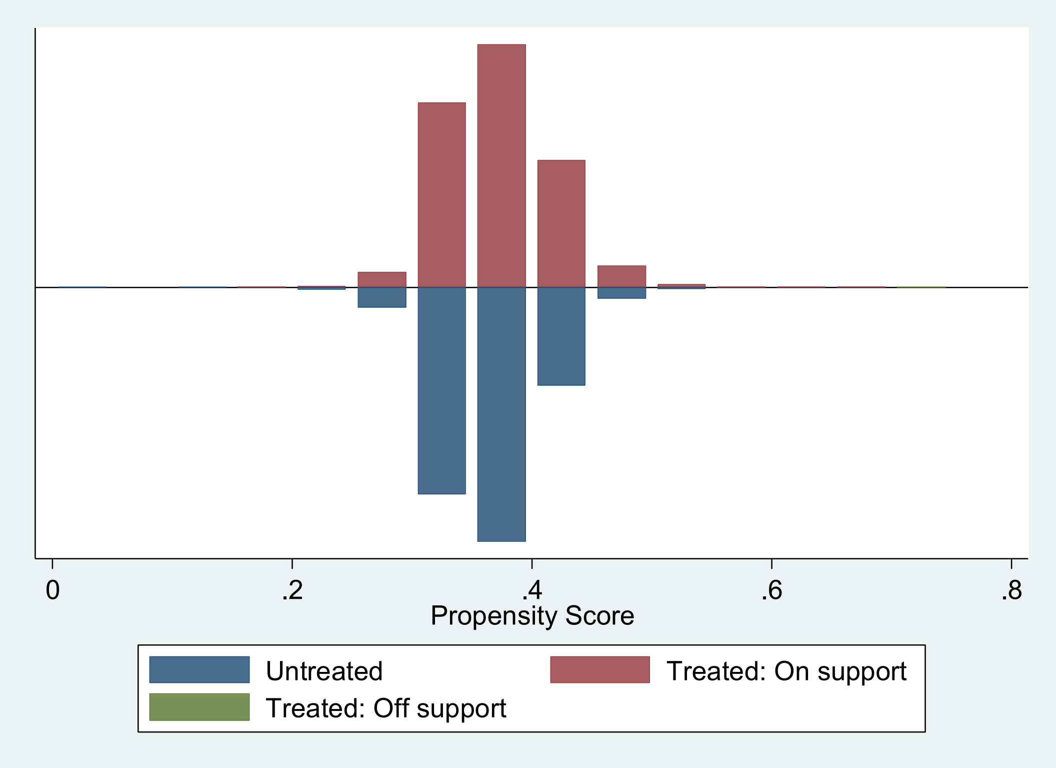
for assessing differences than t-statistics. See Imbens and Wooldridge (2009), who suggest using a normalized difference equal to the difference in averages between the two groups divided by the square root of the sum of the variances for the two groups. This metric has the advantage that larger sample sizes do lead to larger expected t-statistics for the same level of bias. The disadvantage to this metric is that, purely as a practical matter, it is much less widely used than a t-statistic and so many audiences may prefer t-statistics.

Table 3-5: Distributions and Means of Local Capacity Area, Usage and CARE Status for SmartRate Customers, Control Customers and the Residential Population

Characteristic	SmartRate Population	Matched Control Group	t	p
Greater Bay Area	27%	27%	1.17	0.24
Greater Fresno	16%	16%	-0.01	0.99
Kern	27%	27%	-0.44	0.66
Sierra	14%	14%	-1.80	0.07
Stockton	6%	6%	2.02	0.04
Other	9%	10%	-0.63	0.53
June 2011 kWh	539	563	7.30	0.00
July 2011 kWh	809	836	5.20	0.00
Non-CARE	53%	51%	3.95	0.00
CARE	47%	49%	3.95	0.00

Table 3-5 does not address the possibility that although mean values may be similar across groups, the overall joint distribution of these values could differ significantly across the groups. If the table of mean values, as in Table 3-5 shows few differences, then this is very unlikely to be an issue in RCTs or REDs. However, it could be a problem with propensity-score matched control groups. This can be partially addressed by showing histograms of the propensity score for both the treatment group and for a matched control group. An example of this is shown in Figure 3-2, which is taken from the same analysis as Table 3-5. In this case, propensity score histograms are very similar in each group, indicating a well-matched control group.

Figure 3-2: Propensity Score Match Histograms



More important than demonstrating that the groups are comparable across demographic characteristics is demonstrating that they have similar usage patterns. The fundamental assumption underlying the reference load model for each of these analyses is that the control, non-encouraged or matched control group load is an accurate estimate of what the load in the treatment or encouraged group would have been but for being exposed to treatment. The best way to demonstrate that this is a realistic assumption is to demonstrate that loads between the groups are very similar prior to the treatment, under event-like conditions. Showing graphs of hourly average usage in each group for the same set of pre-treatment days used in the regression model can demonstrate this.

An example of these graphs is shown in Figures 3-3 and 3-4 below, which were taken from the same analysis as Table 3-4. The graphs compare usage between the treatment group and the matched control group over a set of days. In this case, no pre-treatment data was available but the rate underlying the CPP rate was the default residential rate, so it was assumed that treatment and control customers would have similar behavior on non-event days. That assumption should only be used if no pre-treatment data is available. For non-event based programs, that assumption is not even an option. In this case, the graphs show that the load is generally quite similar between the groups, but that it does deviate noticeably between groups at some times. Simply seeing the scale and frequency of the deviations can be useful in interpreting the accuracy of the regression output.

Figure 3-3:
Average Usage for SmartRate Population and Control Group
Hot, Non-event Days versus Average Temperature

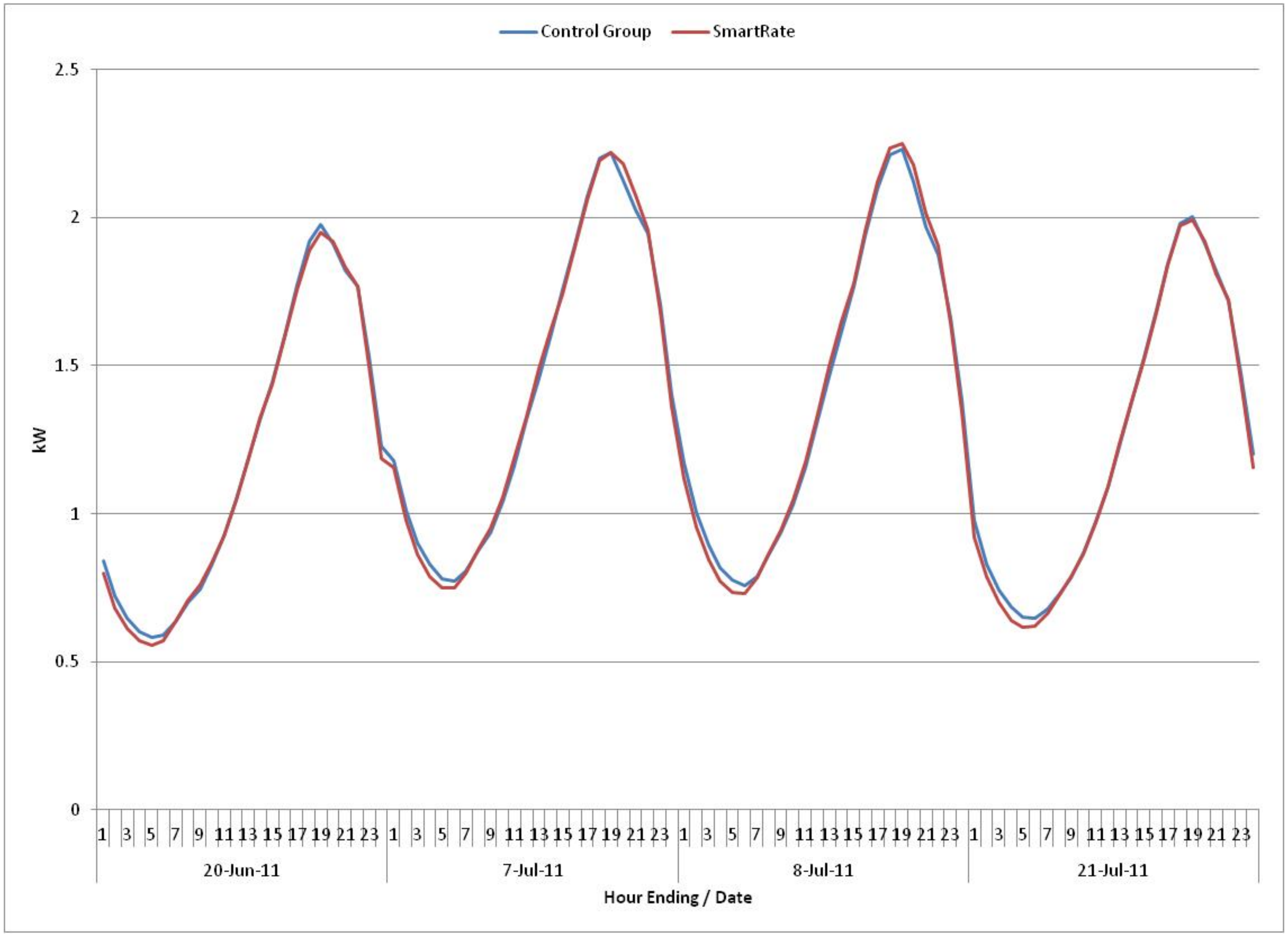
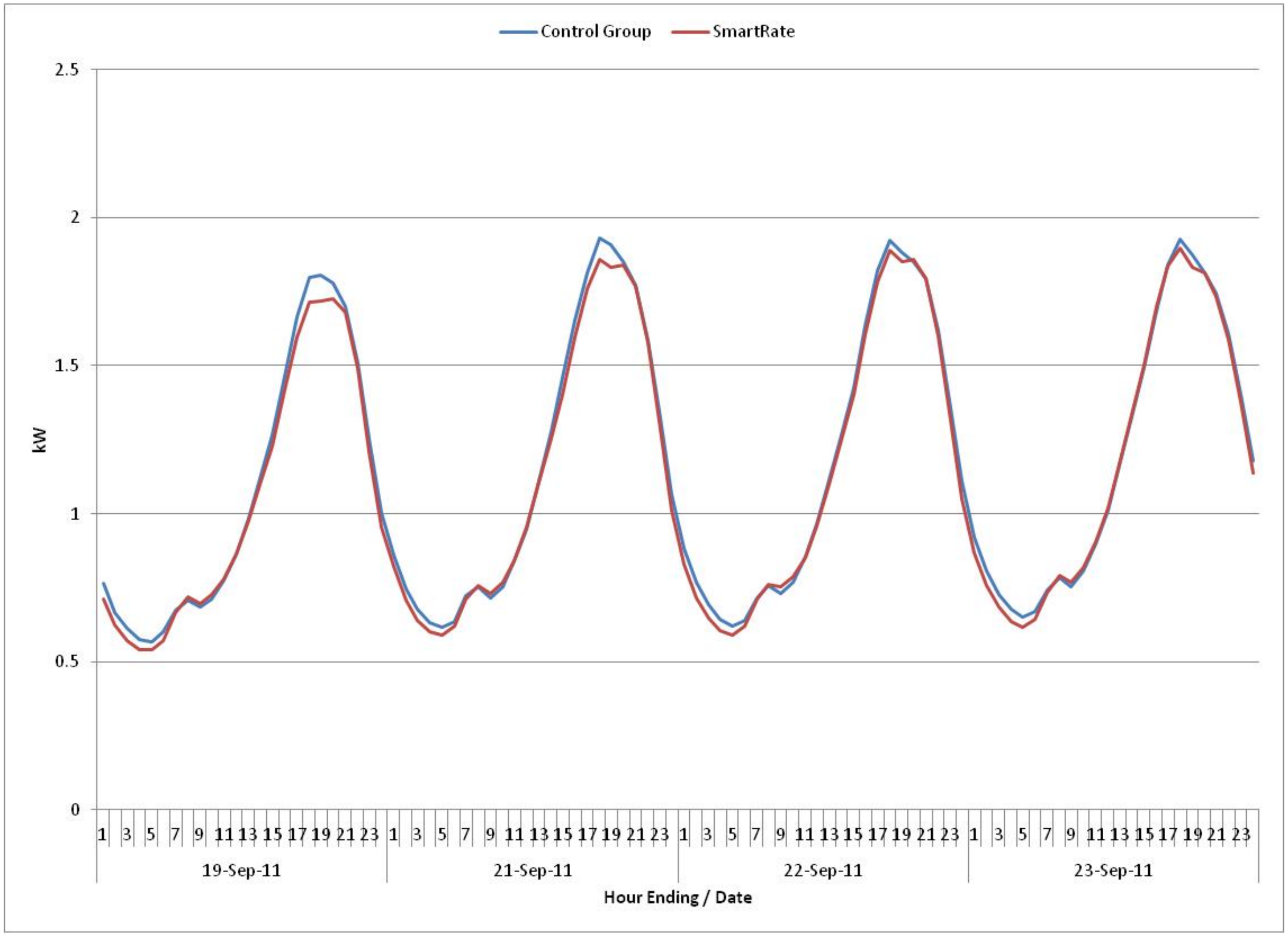


Figure 3-4:
Average Usage for SmartRate Population and Control Group
Hot, Non-event Days versus Average Temperature



To adapt Figures 3-3 and 3-4 to the case of a non-event based treatment, it might be convenient to show the average over different types of days, rather than every day individually. Unlike in the event-based case, for non-event based treatments we are interested in group comparability under many different conditions. A good way to organize the Figures might be according to categories of high temperature on weekdays and weekends. For example, an analog to Figures 3-3 and 3-4 could show average hourly usage in each group on weekdays with high temperatures in the 70s, 80s, 90s and above 100, each as its own daily average, and similarly, but separately for weekends.